Models of speech perception

Paper 9
Foundations of Speech Communication
Lent: Week 2

Katharine Barden
Models of speech perception

- Abstractionist vs exemplar approaches
- Prelexical units
- Modality of representations
- Balance of “top-down” vs “bottom-up” processing
- “Speech is special”
Abstractionist approach

- **Focus on identifying phonological or lexical units**
- **Motivation:**
  - perception seems categorical
  - small set of linguistic units generates infinite set of utterances
  - phonological rules operating on linguistic units
- **Impose order by early abstraction into units of formal linguistics**
- ‘Essence’ + variation
  - preserve the essence (phonemes or features)
  - discard the variation

Phonetic detail is not retained beyond the earliest stages of perception
Motor Theory (Liberman et al., 1967)

• No simple relationship between the acoustic signal and the perceived phonemes
  
  → abandon acoustic invariance

• Invariant units = the speaker’s phonemic gestures

• Gesture = a class of movements by one or more articulators
  
  – e.g. movements for the “voiced labial stop gesture” /b/
    
    • closing and opening of the lips
    • closing the velum
    • starting vocal fold vibration at same time as lips open
Motor Theory

Motivation:

• Different acoustic properties can be the consequence of invariant gestures

  e.g. /d/ rising F2 transitions in /di/
  falling F2 transitions in /du/

  But both of these acoustic patterns result from an alveolar closure gesture that is followed by movement to appropriate tongue position for the vowel.
Motor Theory

Motivation:

• Coarticulation

  – produces variability so that the acoustic signal itself contain no invariant properties

  – the underlying phoneme gestures can be retrieved by compensating for the coarticulatory effects

  e.g. in the context of lip-rounding, listeners will perceptually compensate for the resulting lower of frequencies in surrounding sounds
Motor Theory

Motivation:

• **Categorical perception**
  • Sharp category boundary
  • Poor discrimination of acoustic differences within categories
  • Good discrimination of acoustic differences across categories

  – When a sound is heard, gestures are identified; Sounds that could come from a single type of articulation are responded to as functionally equivalent (e.g. as /b/ or /d/)
Motor Theory

A unifying explanation?

• Speech production and perception share a common representation of phonemes as motor gestures

How does it work...?

• Speech perception is automatically mediated by an innate, specialised speech module to which we have no conscious access
Motor Theory

Evaluation

• Are gestures any less variable than the acoustic signal?
  – Probably not...
  – The Revised Motor Theory (Liberman and Mattingly, 1985) addresses evidence that gestures are variable by suggesting that it is the intended gestures that are perceived rather than the actual movements a speaker makes.
Motor Theory

Evaluation

• How does the transformation from the acoustic signal to the motor gesture happen?
  – Motor Theory postulates an innate module for processing speech
    • How could it be tested?
  – How would this innate module derive invariant gestures from a variable acoustic signal?
Motor Theory

Evaluation

• Evidence that perception precedes production in native language acquisition could call into question the idea that innate gestures underlie production
Motor Theory

Evaluation

• Is the model’s output (a string of abstract intended phonemic gestures) able to account for all influences on speech perception?
  • Internal category structure
  • Talker-familiarity and perceptual learning effects
  • Systematic phonetic detail
  • Top-down contextual influences
  • Error correction
Motor Theory

Evaluation

• Are there alternative ways of overcoming the ‘problem of invariance’?
  – E.g. more complex relationships between the acoustic signal and linguistic units
Acoustic Invariance

• Alternative response to ‘problem of invariance’
  – From static templates to relational invariants
  – Lexical Access From Features (LAFF – Stevens, 1980s)
    • Change the unit - search for invariant acoustic or auditory correlates of phonological features (rather than phonemes)
Lexical Access From Features
(Stevens, 1980s)

• Each word in the mental lexicon is a **matrix of features** which are distributed over time to mirror the way each appears in time

• Features that can change a lot due to coarticulatory effects are marked as **modifiable** (M)
**Lexical Access From Features**

(Stevens, 1980s)

<table>
<thead>
<tr>
<th>Type of feature</th>
<th>Specific feature</th>
<th>t</th>
<th>æ</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Landmark</strong> <em>(manner of constriction: articulator-free)</em></td>
<td>consonantal</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>sonorant</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>continuant</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>syllabic</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>lateral</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>strident</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Place of constriction</strong> <em>(articulator-bound)</em></td>
<td>lips round</td>
<td>+</td>
<td>-</td>
<td>+M</td>
</tr>
<tr>
<td></td>
<td>tongue blade</td>
<td>+</td>
<td>-</td>
<td>+M</td>
</tr>
<tr>
<td></td>
<td>anterior</td>
<td>+</td>
<td>-</td>
<td>+M</td>
</tr>
<tr>
<td></td>
<td>distributed</td>
<td>+</td>
<td>+</td>
<td>-M</td>
</tr>
<tr>
<td></td>
<td>tongue dorsum</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>back</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td><strong>Other articulators</strong></td>
<td>voiced</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>nasal</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

General form of the lexical representation of the word *tan* according to LAFF
Lexical Access From Features

Evaluation

• Difficult to define the acoustic properties of some features

• Does not include intrinsic duration of features (just their sequencing)
Lexical Access From Features

Evaluation

- M’ feature → some allowance for variation
  But does it take into account the role of context in determining acceptable variation?
  - Assimilation across a word boundary
    e.g. *tan badly*
  - the place of constriction may be labelled as M (modifiable)
  - BUT the bilabial PoA is only acceptable in the appropriate conditioning phonological context i.e. before a bilabial consonant (not before an alveolar consonant)
  - M feature allows variation but does not specify appropriate context for such variation
Lexical Access From Features

Evaluation

• Output is features rather than phonemes
⇒ greater detail can influence perception

• But there are still some aspects of phonetic detail that cannot be captured in a featural representation e.g. mistakes vs mistimes
Lexical Access From Features

Evaluation

• Output = distinctive features mapped onto words

• How can this account for...
  – long-domain effects?
  – talker-familiarity effects?
  – systematic phonetic detail relating to structures higher than the word?
  – top-down influences?
Lexical Access From Features

Evaluation

• (How) is perception related to production?
Exemplar/episodic theories

Attempt to account for evidence that listeners remember details relating to specific episodes

- Listeners are more accurate at recognizing that they’ve heard a word before if it is repeated by the same talker and at the same speaking rate (Bradlow et al., 1999)
- Facilitatory effect of familiarity with a talker (see last week’s handout)

⇒ retention of surface form

- Cannot be captured by an abstractionist approach where phonetic information is discarded early in processing
Exemplar theories

• Each stimulus (e.g. a word) leaves a unique trace in memory
• When a new stimulus is presented, these memory traces are activated in proportion to their similarity to the stimulus
• Conscious experience of perception occurs as a result of the combined activation of these previously stored exemplars
Exemplar theories

- No explicit prelexical code
- The general similarity between many exemplars ⇒ **Emergent categories**
- Categories are **self-organising** and can be of many sizes ⇒ Listeners can make use of systematic phonetic variation that reflects linguistic structure of many types
- Close similarity between few exemplars ⇒ **Talker-specific details can influence perception**
Exemplar theories

• Can account for plasticity
  – categories are continually evolving as new experiences create new memory traces
Exemplar theories

- **Top-down effects**
  
  e.g. Ganong Effect:
  
  - memory traces exist for the real word but not for the nonsense word
  
  - therefore activation of real-word memory traces will contribute to the percept
  
  - whether it is sufficient to alter the percept will depend on the similarity of the stimulus to the traces
Multimodal exemplar theories

• Multimodal memory traces…
• Experience of one modality can activate all/some of others through associated memories
• Meaning is achieved through combined activation of relevant multimodal memory traces on exposure to a stimulus
• Can account for multimodal effects on perception e.g. McGurk effect, gestures
• A more holistic approach to communication – from sound to meaning
Exemplar theories

Problems...?

• What is an exemplar?
  – What are the temporal boundaries of an ‘episode’? What size/s ‘units’ are compared?
  – At what level/s of processing are exemplars stored?
  – How much of the context is stored?

• What is similarity? What counts as ‘similar’?
Exemplar theories

Problems...?

• Plasticity vs stability

• How to test... Implementations are necessarily limited in comparison to human experience

• Neurologically plausible? Memory capacity?
Exemplar theories

Within the general exemplar approach, there are various different models...

• Reading on exemplar approach:
  – David Pisoni, Keith Johnson, Stephen Goldinger

Models

• Goldinger: MINERVA – a computational implementation of a very limited exemplar model
• Jusczyk: WRAPSA – based on child language acq
• (?) Kuhl: Native Language Magnet Model (can be viewed as exemplar-based?)
• S. Hawkins and Smith: Polysp
• J. Hawkins. Memory-prediction framework – not specifically about speech
TRACE: a connectionist model
(McClelland and Elman, 1986)

Aimed to:

– identify single words

– account for categorical perception, Ganong effect and other traditional phonetic findings that were considered important in 1970s-1980s
TRACE: a connectionist model
(McClelland and Elman, 1986)

• TRACE is a good example of an interactive model
  (making use of bottom-up and top-down information)

• Probabilistic: Several words can be “activated” simultaneously, and the word that is identified is the most probable one (the most highly activated)
Structure of TRACE

3 layers:
- words
  - cat
  - deep
  - dog
  - dark
  - seep
  - lark
- phonemes
  - b
  - d
  - s
  - l
  - o
- features
  - voice
  - bilab
  - alv
  - cons
  - voc

INPUT: featural information, one time-slice at a time

OUTPUT: one word (or probabilities for several words)
Structure of TRACE

3 layers:

words

cat
dog
dark
seep
lark

phonemes

b
d
s
l
o

features

voice
bilab
alv
cons
voc

Each unit connects directly with every unit in its own layer, and in the adjacent layer(s).
Connections BETWEEN layers excite (activate) units in other layers.
Connections WITHIN a layer inhibit (lower activation of) units in same layer.
Processes of TRACE

• Each unit has a resting activation level
• Each connection has a numerical weight that determines how fast it conducts information and thus how much information is passed in a given time
• Output was originally a single word, but modified to be probabilities for different words
TRACE

Evaluation:
• Top-down effects can account for e.g. Ganong Effect, but are sometimes too powerful
  – strong influence of word level on phoneme level can prevent detection of serious mispronunciations
• Surface phonetic variation is discarded
• No higher level structures (incomplete)
Summary

• Diverse approaches to modelling speech perception
• Varied in scope – identification of phonemes, words, or all the way to meaning...
• Assumptions: abstractionist vs exemplar, motor vs acoustic invariants, features vs phonemes, top-down vs bottom-up...
Next week…

• Speech production
  – Models of coarticulation
Reading

Main sources:


Exemplar theories:
Reading

Connectionism/neural networks

TRACE

Further reading on specific models:

Motor theory

WRAPSA
Reading

Specific models:
Native Language Model – expanded.

Polysp

Memory-prediction framework

Direct realism